ChatGPT

MAP – Charting Student Math Misunderstandings competition overview

The **MAP – Charting Student Math Misunderstandings** competition is a featured code competition on Kaggle hosted by **The Learning Agency LLC** with support from Vanderbilt University and the Eedi educational platform. The goal is to build an NLP model that predicts a student's likely misconceptions from their explanation of a multiple-choice math question. When students answer diagnostic questions on Eedi, they sometimes explain their answer. These explanations often reveal misconceptions, but tagging explanations manually is time-consuming. The competition asks participants to create models that can suggest likely misconceptions so teachers can address errors more effectively [1]. The competition opened on **10 July 2025** and runs until **15 October 2025** [2], with a total prize pool of **US$55 000** [3].

Timeline and prizes

Key dates (all deadlines are 23:59 UTC) [2]:

Milestone	Date
Competition opens	**10 Jul 2025**
Entry deadline (no new teams)	**8 Oct 2025**
Team merger deadline	**8 Oct 2025**
Final submission deadline	**15 Oct 2025**

Leaderboard prizes include **$20 000** for 1st place, **$12 000** for 2nd, **$8 000** for 3rd, and **$5 000** for each of 4th–6th places [3].

Dataset and tasks

Data description

The dataset consists of diagnostic questions from Eedi. After choosing a multiple-choice answer, students may provide a written explanation. Each row in the **train/test** CSV represents one student response and contains [4]:

Column	Description
`QuestionId`	Unique identifier for the question
`QuestionText`	Text of the question. OCR has been applied to images so the text is available without processing the images [5].
`MC_Answer`	The multiple-choice answer selected by the student

Column	Description
`StudentExplanation`	Free-text explanation given by the student
`Category` *(train only)*	Relationship between the selected answer and explanation. Possible values are `True_Correct`, `True_Incorrect`, `True_Misconception` (correct answer but explanation shows a misconception) and their `False_*` counterparts [4].
`Misconception` *(train only)*	Specific math misconception tag (e.g., "Incomplete fraction simplification"); `NA` if no misconception applies [4].

The **sample submission** file contains two columns— `row_id` and `predictions`. For each row in the test data, participants must predict up to three `Category:Misconception` combinations separated by spaces. Predictions beyond the top three are ignored [4].

Target tasks

Models must perform three sub-tasks:

1. **Determine if the multiple-choice answer is correct.** A model must decide whether the student's chosen answer is right or wrong (the `Category` label encodes this as `True_*` or `False_*`).
2. **Determine whether the explanation reveals a misconception.** Some explanations show misunderstandings even when the answer is correct (e.g., the student says a fraction cannot be simplified further). This part influences whether the `Category` is labelled `_Correct`, `_Incorrect`, or `_Misconception` [5].
3. **Identify the specific misconception tag** (e.g., "Incomplete"). There is exactly one misconception label per explanation when applicable [6].

Additional data notes

- Questions are typically displayed as images. Organisers used human-in-the-loop OCR to extract question text and provide it in `QuestionText` [5]. The original images and bounding boxes are included for participants who want to work with images [7].
- The dataset license is **MIT** [8].
- The `train.csv` includes **15** unique questions. Kaggle staff member Chris Deotte confirmed via leaderboard probing that the test set does **not** contain any new questions; therefore participants can cross-validate using K-folds without a group split on question ID [9]. [9]
- Participants discovered **12 mis-labelled rows** in the training data (rows with `QuestionId=31778` labelled `True_Correct` when the correct answer is 6). Removing these incorrect labels improved models and is implemented in the "is_correct" feature described below [10].

Evaluation metric (MAP@3)

Submissions are evaluated using **Mean Average Precision @ 3 (MAP@3)**. For each observation, participants may submit up to three predicted `Category:Misconception` pairs ranked by confidence. MAP@3 is defined as [6]:

$$\text{MAP@3} = \frac{1}{U} \sum_{u=1}^{U} \sum_{k=1}^{\min(3,n)} P(u,k) \cdot \text{rel}(u,k),$$

where U is the number of observations; n is the number of predictions; $P(u,k)$ is the precision at cutoff k ; and $\text{rel}(u,k)$ is an indicator equal to 1 if the prediction at rank k is the correct label and 0 otherwise. Once the correct label is predicted, further positions do not contribute. The example in the competition description shows that predicting the correct label anywhere in the top three yields an average precision of 1.0 [6] .

Submission requirements and rules

- **Code competition**: submissions must be Kaggle notebooks with internet disabled and a maximum runtime of **9 hours** on CPU/GPU [11] . External data must be public/free; private data or test-set probing is prohibited.
- The final output file must be named `submission.csv` and contain predictions as described above [4] .
- Models may load pre-trained weights; these must be made available via Kaggle Datasets or provided in the notebook [11] .
- Participants must register for the competition and agree to Kaggle's general competition rules. Private code/data sharing outside of approved teams is not allowed [12] .

Insights from Kaggle discussion forum

Participants have been actively discussing strategies and clarifying dataset issues. Below are summaries of the most commented threads (based on comment counts) and pinned posts.

Welcome and administrative threads

- **"Get started here"** (pinned): Kaggle staff member Maria Cruz welcomes newcomers, provides links on site etiquette, how to enter a competition via Kaggle Notebooks/Models, and encourages use of the Team Up feature [12] . Participants introduced themselves; new users thanked the community and mentioned exploring beginner competitions like Titanic or House Prices. A notable comment asked how to install libraries such as `bitsandbytes` without internet; community members later advised uploading wheel files as a dataset and installing them locally [13] .
- **Official Discord channel** (pinned): Maria Cruz announced an official Kaggle Discord server (discord.gg/kaggle) for general discussion. She reminded competitors that Discord channels are public and not monitored by Kaggle staff; important questions should still be posted on the forum. Sharing private code or data outside of team channels is prohibited [14] .

Model-building threads

- **"From 1.5B to 7B: My Best Single Model LB Score So Far"** – Jaydev Tonde described experiments fine-tuning various large language models (LLMs) using **QLoRA**. His models (DeepSeekMath-7B, Mathstral-7B, Qwen3-8B, etc.) achieved validation MAP@3 around **0.939–0.943** and public leaderboard scores around **0.940–0.943**. He trained using an 80/20 split and said the validation set was random [15] . Commenters asked about ensembling; Jaydev noted he had not tried ensembles yet but planned to. Others asked whether he trained full model weights; he clarified that he used **quantized LoRA** (QLoRA) rather than full-parameter fine-tuning [15] .

- **"Q-LoRA or just LoRA for training ?"** – Participant *Alan Turing* asked why his QLoRA-fine-tuned Gemma2–2B model achieved CV ≈ 0.93 but leaderboard MAP@3 ≈ 0.843. Another user responded that quantization itself should not cause such a large drop and that a properly tuned 4-bit QLoRA retains near-full performance [16]. Alan shared his notebook for feedback; the conversation concluded that inference details, not LoRA vs. QLoRA, likely caused the discrepancy.

Infrastructure threads

- **"Internet off submission"** – New competitor Roushan Kumar asked how to install libraries like `bitsandbytes` when notebooks must run offline. Experienced users suggested downloading wheel (`.whl`) files locally or from Kaggle and uploading them as a dataset; these can then be installed in the offline submission environment [13].
- **"Why Notebook Timeout"** – A user noted that their notebook completed in under 10 minutes locally but timed out at 9 hours during submission. Chris Deotte explained that local notebooks only run inference on **3** test samples, whereas final submissions run on **about 16 000** test rows. To estimate runtime, he recommended inferring on a smaller subset (e.g., 1 000 samples) locally and scaling up [17].
- **"Val loss = NaN with QLoRA"** – A competitor encountered NaN validation loss when training a 14-B QLoRA model. Chris Deotte advised evaluating the model before training to ensure the data and evaluation metric are functioning; he also suggested adjusting the learning rate, using gradient clipping or changing precision (e.g., bf16/fp32) [18].

Data quality and probing threads

- **"Does Test Data Have Questions Different Than Train Data?"** – This thread became the most commented (≈50 comments). Chris Deotte observed that the train set contains only **15** unique questions and asked whether the full test set contains new questions. Participants used **leaderboard probing** (writing code that outputs a diagnostic value instead of predictions) and confirmed that **no new questions** appear in the test set [9]. Because the test and train sets share the same questions, Chris proposed a powerful feature: create an `is_correct` flag that identifies whether the student's multiple-choice answer matches the correct answer for that question. By constructing a dictionary of question-answer pairs, he improved cross-validation MAP@3 from **0.95** to **0.957** and public leaderboard scores accordingly [10]. He also noted **12 mis-labelled rows** (QuestionId 31778 with wrong correct answer) and removed them when building the `is_correct` feature [10].

The conversation also discussed cross-validation strategies: since there are no new questions in the test set, participants do not need a group K-fold by question and can use plain K-fold or stratified K-fold [9]. Some users raised concerns about label noise and whether similar mis-labelling might occur in the hidden test data. Others asked how leaderboard probing works; Chris explained that a notebook can submit a constant score (e.g., 0.940) and examine the log outputs to infer information about the test set [9]. Overall the thread highlighted that data quality and simple engineered features (like `is_correct`) are critical.

- **"Is Public LB 0.950 a Resistance Level?"** – Another popular thread (≈18 comments) in which Chris Deotte noticed that public leaderboard scores seemed capped at **0.950**. He speculated that errors and label disagreements in `test.csv` might limit the maximum score. Many participants shared their cross-validation (CV) and leaderboard scores, with CV MAP@3 values around **0.952–0.953** and leaderboard scores near **0.948–0.951** [19]. Some believed the barrier could be surpassed (e.g., 0.955+) [20], while others noted that ensembling or data cleaning might yield marginal improvements. Discussions highlighted that robust models and cleaning mis-labelled rows are likely required; some suggested training separate models per question ID

or checking for unseen `Category:Misconception` combinations [21]. The consensus was that slight improvements above 0.950 are possible but that label noise sets an upper bound.

- **"Same Student Explanation with different MC Answer selected and different Misconception labels assigned"** – A participant observed that identical student explanations in `train.csv` sometimes received different `Category` and `Misconception` labels depending on the selected multiple-choice answer. For example, the explanation "Because there are 9 triangles and 3 of them are not shaded" was labelled `True_Correct` when paired with the simplified answer $\frac{1}{3}$ but `False_Misconception` **(Incomplete)** when paired with $\frac{3}{9}$ [22]. Commenters explained that the difference arises because $\frac{3}{9}$ is not simplified; raters treated the answer as "incomplete" but correct reasoning when simplified [23]. This thread reinforced that the `Misconception` label depends on both the explanation and the exact form of the answer.

- **"I think there are errors in the labels as well"** – This post shared another example where the student's reasoning suggested a misconception was mislabeled. The author pointed out that the student's explanation ("1 3rd is half of 3 6th") reveals they divided numbers incorrectly, yet the row was labelled `True_Neither`. The post emphasised that understanding students' reasoning is more important than whether an answer is right or wrong [24]. No responses were recorded.

Miscellaneous threads

- **"Are Kaggle servers down?"** – A brief post noting temporary submission issues; there were no substantial comments.
- **"Relatively New to ML/Coding"** – A newcomer asked for advice on getting started. At the time of viewing, no one had responded [25].

Takeaways

- **Data quality and simple features matter**: The `is_correct` feature—flagging whether a student chose the right answer—dramatically boosts performance when the train and test sets contain the same questions. Identifying and correcting mis-labelled rows also helps.

- **Label noise may cap performance**: Top competitors report CV scores around 0.952–0.953 and LB scores around 0.949–0.951, suggesting that errors in the dataset and disagreements among human raters may limit the maximum achievable MAP@3 [19].

- **Large models vs. efficient fine-tuning**: Participants experiment with fine-tuning LLMs (e.g., Qwen, DeepSeekMath) using LoRA/QLoRA. Results show that carefully tuned LoRA models can achieve high scores, but quantization alone does not harm performance. Ensemble strategies and multi-model approaches remain areas for exploration.

- **Community support**: Kaggle's discussion forums and official Discord provide resources for newcomers and a space to troubleshoot issues such as offline package installation or notebook timeouts [13] [17].

This report summarises the competition structure and synthesises insights from the most active discussion threads up to **14 August 2025**. Future developments may occur as competitors continue to refine models before the final submission deadline.

---

1 MAP - Charting Student Math Misunderstandings | Kaggle
https://www.kaggle.com/competitions/map-charting-student-math-misunderstandings

2 11 MAP - Charting Student Math Misunderstandings | Kaggle
https://www.kaggle.com/competitions/map-charting-student-math-misunderstandings/overview/timeline

3 MAP - Charting Student Math Misunderstandings | Kaggle
https://www.kaggle.com/competitions/map-charting-student-math-misunderstandings/overview/prizes

4 5 7 8 MAP - Charting Student Math Misunderstandings | Kaggle
https://www.kaggle.com/competitions/map-charting-student-math-misunderstandings/data

6 MAP - Charting Student Math Misunderstandings | Kaggle
https://www.kaggle.com/competitions/map-charting-student-math-misunderstandings/overview/evaluation

9 10 MAP - Charting Student Math Misunderstandings | Kaggle
https://www.kaggle.com/competitions/map-charting-student-math-misunderstandings/discussion/589400

12 MAP - Charting Student Math Misunderstandings | Kaggle
https://www.kaggle.com/competitions/map-charting-student-math-misunderstandings/discussion/589185

13 MAP - Charting Student Math Misunderstandings | Kaggle
https://www.kaggle.com/competitions/map-charting-student-math-misunderstandings/discussion/599008

14 MAP - Charting Student Math Misunderstandings | Kaggle
https://www.kaggle.com/competitions/map-charting-student-math-misunderstandings/discussion/589184

15 MAP - Charting Student Math Misunderstandings | Kaggle
https://www.kaggle.com/competitions/map-charting-student-math-misunderstandings/discussion/598250

16 MAP - Charting Student Math Misunderstandings | Kaggle
https://www.kaggle.com/competitions/map-charting-student-math-misunderstandings/discussion/598164

17 MAP - Charting Student Math Misunderstandings | Kaggle
https://www.kaggle.com/competitions/map-charting-student-math-misunderstandings/discussion/599105

18 MAP - Charting Student Math Misunderstandings | Kaggle
https://www.kaggle.com/competitions/map-charting-student-math-misunderstandings/discussion/598868

19 20 21 MAP - Charting Student Math Misunderstandings | Kaggle
https://www.kaggle.com/competitions/map-charting-student-math-misunderstandings/discussion/591054%233266231

22 23 MAP - Charting Student Math Misunderstandings | Kaggle
https://www.kaggle.com/competitions/map-charting-student-math-misunderstandings/discussion/598184

24 MAP - Charting Student Math Misunderstandings | Kaggle
https://www.kaggle.com/competitions/map-charting-student-math-misunderstandings/discussion/597944

25 MAP - Charting Student Math Misunderstandings | Kaggle
https://www.kaggle.com/competitions/map-charting-student-math-misunderstandings/discussion/599040